Just AI

On the way to industrial NLP-platform:

transformers, microservices, architecture

Murat Apishev

NLP Team Lead, Just AI

Contents

- New classifier for our chatbot platform
- Paraphrase for user inspiration
- Problems with old NLP service
- The new ML platform

Models for products

Classification is everything

Just AI

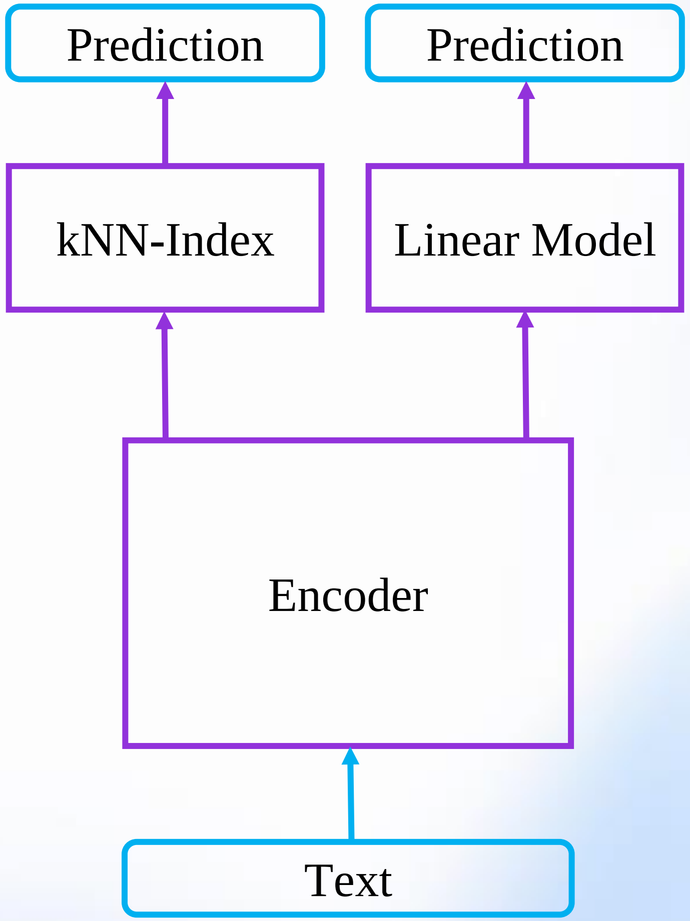
- There are many difficulties and they are known:
  - Closely related classes
  - Imbalanced classes
  - Too few samples
- Problems are solvable for DS
- And are NOT solvable for our regular users
- Classifiers should always work well out of the box

Transformer for platform

Just AI

Transfer Learning is a good all-round solution for complex cases

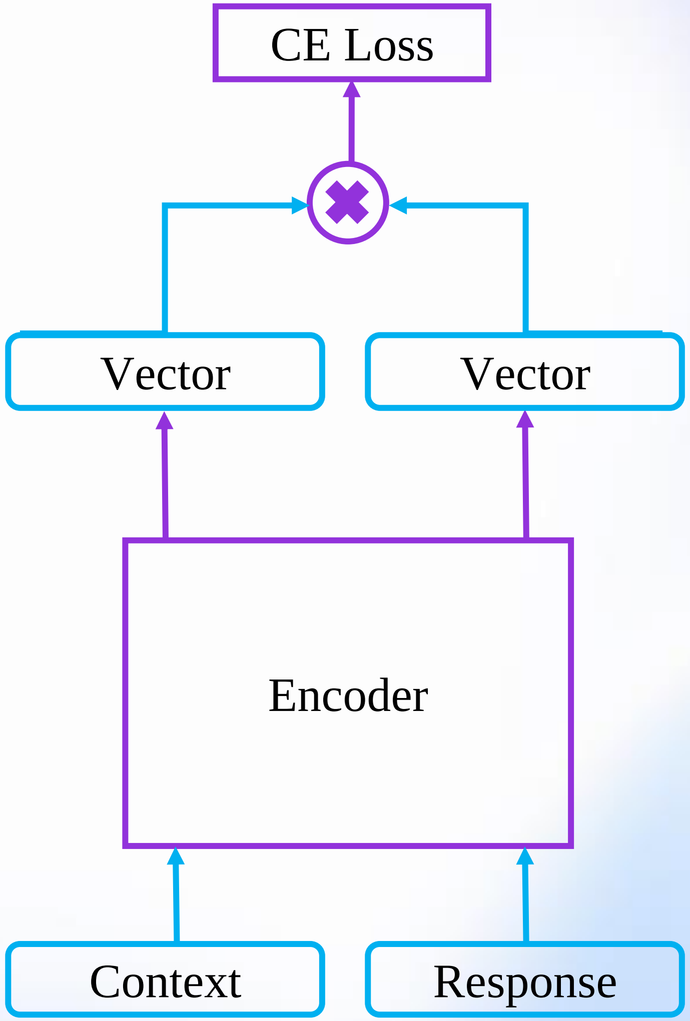
- Use pre-trained BERT-like models with better understanding of the meaning of the text
- Add lightweight trainable head
- Get a good classifier working out of the box



Fine-tuning

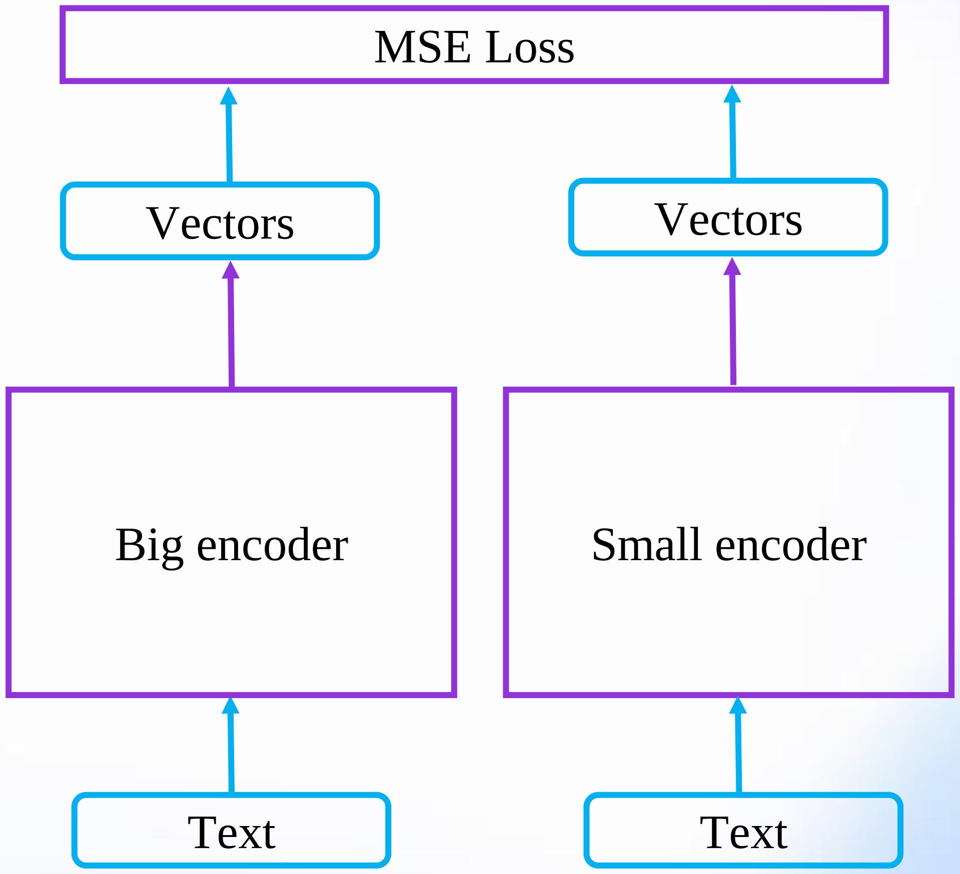
Just AI

- Internal dialog dataset (> 1M dialogs)
- **Proxy task**: increase similarity between a phrase and its context
- **Result**: +5% on average to the accuracy of the best model with a linear head



Acceleration

- **Distillation + ONNX**:
  - CPU speedup: x4.5
  - GPU speedup: x1.7
  - Average accuracy loss: < 1%



Benchmarks

Just AI

- 10 datasets in Russian from different chatbots subject areas

- Typical dataset example:
  - 50 classes
  - 430 train samples
  - 290 test samples

- 3 / 10 datasets are public:
  - HWU-20 Ru *
  - Chatbots-Ru *
  - Russian Intents Dataset **

* https://github.com/AutoFAQ/Intent-Recognition-SaaS-Evaluation
** https://www.kaggle.com/datasets/constantinwerner/qa-intents-dataset-university-domain

Results

Just AI

- What is being compared:
  - New Transformer classifier (inside our platform)
  - Updated old Classic ML (log-regression) and DL (CNN) algorithms (inside our platform)
  - Dialogflow classifier (inside its platform)

- Average results for 10 datasets:

Classifier	F1-micro	F1-macro
Classic ML	0.768	0.704
DL	0.794	0.741
Transformer	**0.841**	**0.803**
DIalogflow	0.785	0.745

Results for public datasets

Just AI

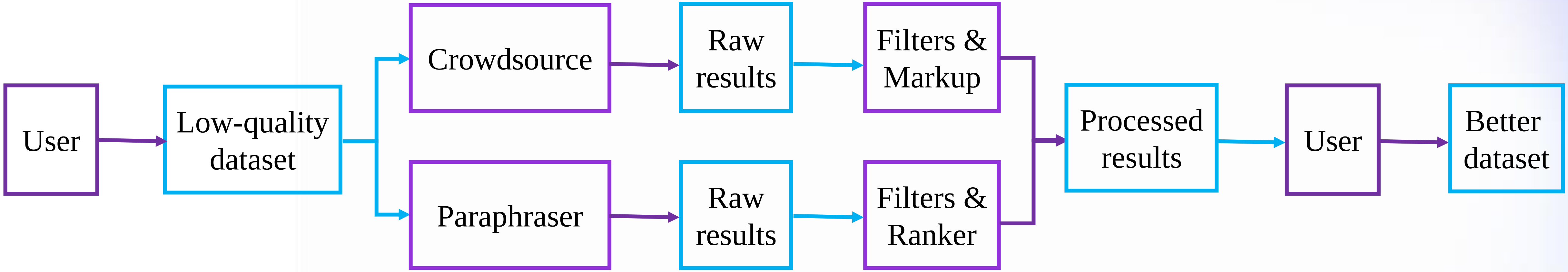
Dataset	Classifier	F1-micro	F1-macro
HWU-20 Ru • Classes: 20 • Train: 100 • Test: 100	Classic ML	0.67	0.65
	DL	0.80	0.79
	Transformer	**0.92**	**0.92**
	DIalogflow	0.72	0.72
Chatbots-Ru • Classes: 79 • Train: 5.6K • Test: 1.4K	Classic ML	0.73	0.71
	DL	0.78	0.77
	Transformer	**0.85**	**0.83**
	DIalogflow	0.63	0.62
RID • Classes: 142 • Train: 13K • Test: 900	Classic ML	0.95	0.94
	DL	0.93	0.93
	Transformer	**0.95**	**0.95**
	DIalogflow	0.94	0.94

Additional help for the user

Just AI

- It happens that there are no training data and logs to create them
- Sometimes users try to compose train phrases, it is a slow process
- The resulting samples have problems:
  - small dataset size
  - poor vocabulary
- Dataset expansion and inspiration options:
  - Crowdsourcing
  - Paraphrasing

Product scenarios



	Crowdsourcing	Paraphrasing
Dataset size	~10^2	1
Duration	Tens of minutes	Seconds
Price	High	Low
Quality and diversity	Good	Customizable
Tasks complexity	Higher	Lower
Solution	Toloka (RU)	T5-based model

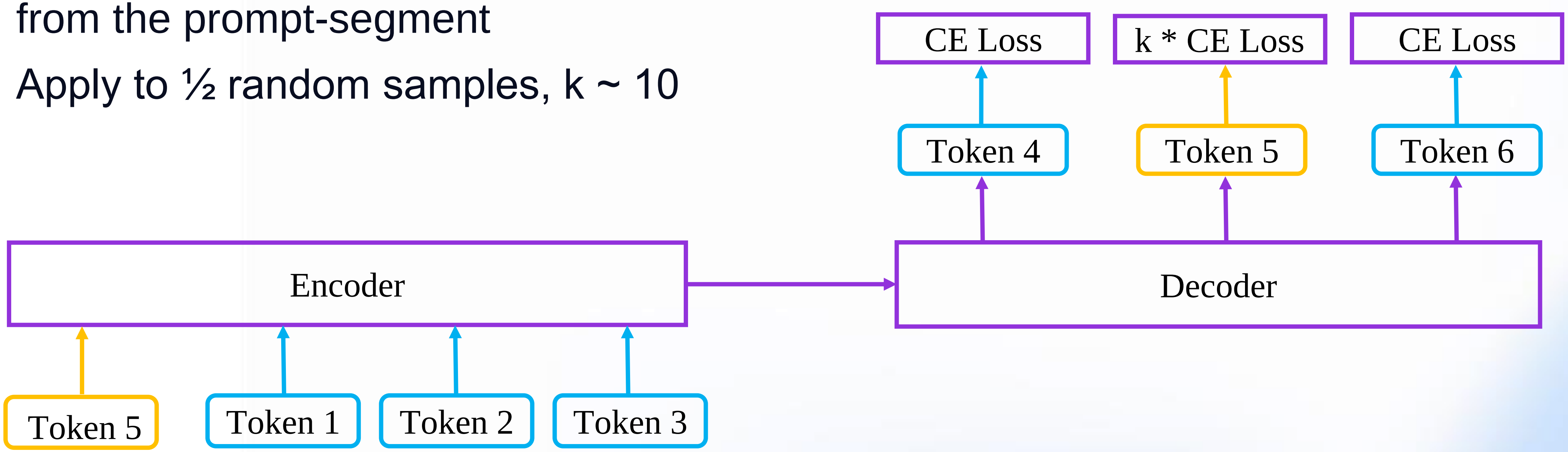
Paraphrase model

Just AI

- Sequence-to-sequense task
  - sberbank-ai/ruT5-base (Russian)
  - t5-base (others)
- Open datasets + backtranslation of open and company internal data
  - > 2M phrase pairs (per language)
- Encourage specific substrings to appear in the result (to use with NER)
  - Special attention to dates
- Different generation parameters presets

Training with text generation control

- Commutative samples
- Hint model with a prompt-segment
- Multiply the penalty for sequences from the prompt-segment
- Apply to ½ random samples, k ~ 10



Examples

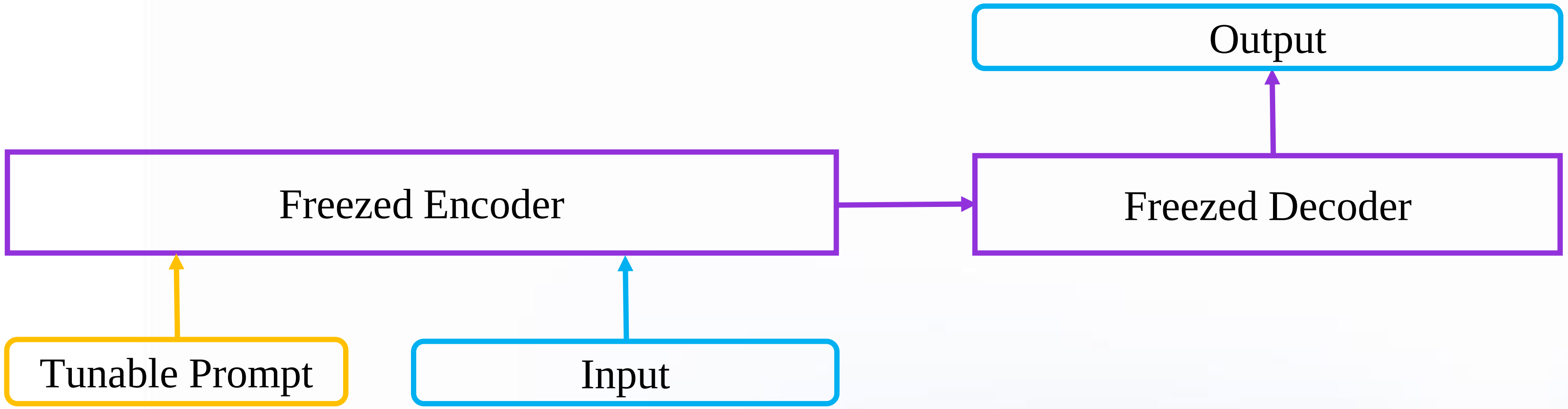
Just AI

Original phrase	Add	Remove *	New phrase
За окном снег, а значит скоро новый год, ура!	-	-	За окошком снег и, следовательно, скоро Новый Год! Ура!
Санкт-Петербург назван в честь святого Петра	-	-	Петербург носит имя Святого Петра.
Хочу закрыть счет в вашем банке	аккаунт	хочу	Я хотел бы закрыть свой аккаунт у вас в банке.
Курьер сообщил, что у него нет части заказа, я хочу вернуть деньги.	монеты	-	Курьер сказал, что части заказов у него не было, мне монеты хочется вернуть.
Курьер сообщил, что у него нет части заказа, я хочу вернуть деньги.	новый год	заказ, заказа	Курьер в новый год сказал, что не имеет части ордера, хочу возврата.

* removing is implemented through bad_word_ids in transformers

Add style options via prompt-tuning

- New model capabilities without retraining
- **Example:** text simplification
- RuSimpleSentEval dataset *
- Trained model Prompt-tuning (ru-prompts library **)



* https://github.com/dialogue-evaluation/RuSimpleSentEval
** https://github.com/ai-forever/ru-prompts

Examples

Just AI

Original phrase	New phrase with simplification
Пётр I Алексеевич, прозванный Великим (9 июня 1672 года — 8 февраля 1725 года) — последний царь всея Руси (с 1682 года) и первый Император Всероссийский (с 1721 года).	Пётр I Алексевич был последним царем Руси с 1682 и первым Императором Всея Руси с 1721
Пушкин один из самых авторитетных литературных деятелей первой трети XIX века.	Пушкин - авторитетный литературный деятель первой трети 19 века.
В природном очаге заражение обычно происходит через укус блохи, ранее питавшейся на больном грызуне.	Заражение происходит из-за укуса блохи.

Evaluating models

Just AI

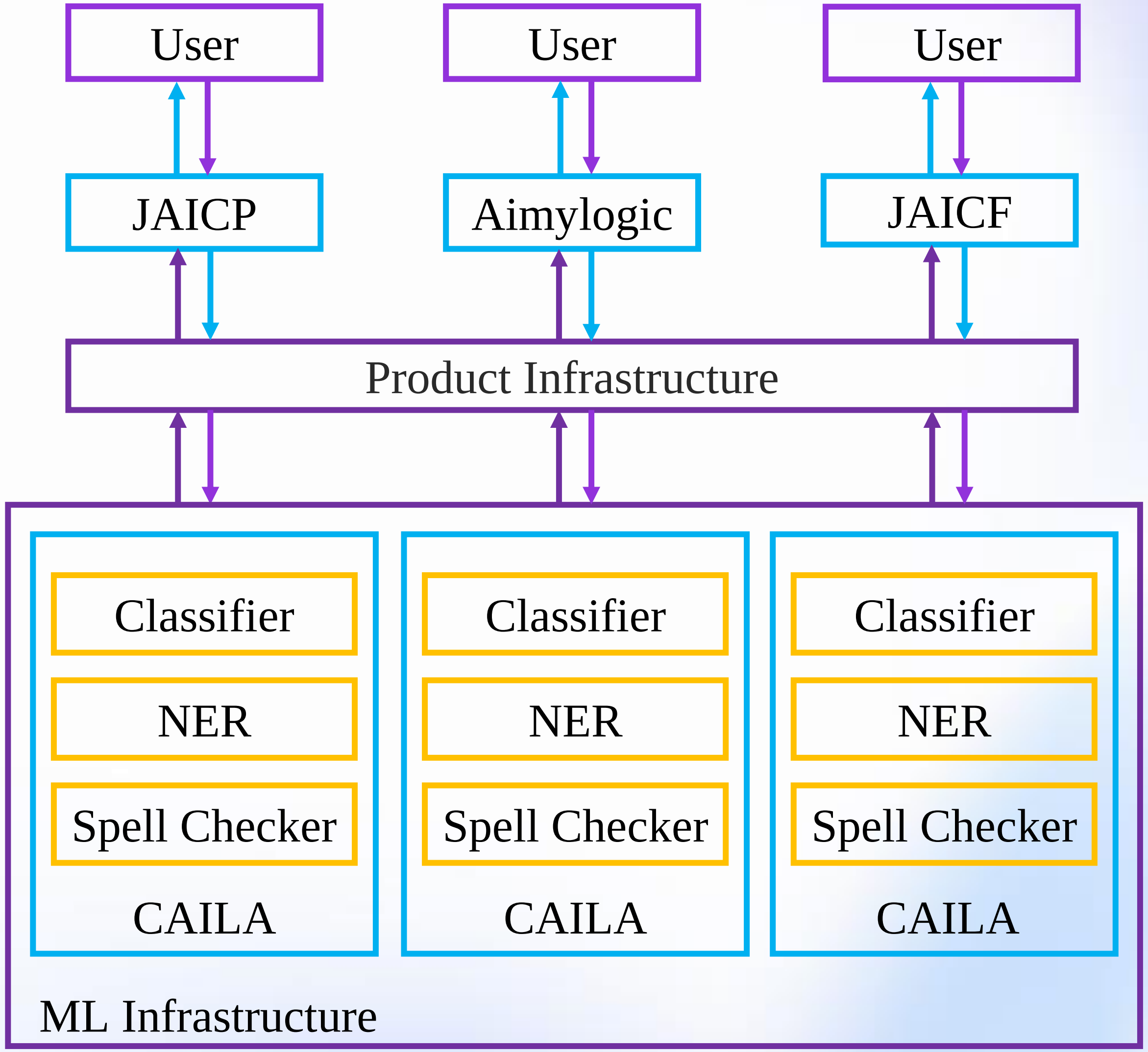
- Several assessors evaluate on general and subject phrases
- Correctness score (200 samples) — 70% for the best model:
  - Meaning preservation
  - Grammar and incorrect words
  - Diversity
- Required words appearance score (170 samples) — 80% for best model
- Average simplification length reduction (75 samples) — 25%

Platform for models

Where to integrate new models

Just AI

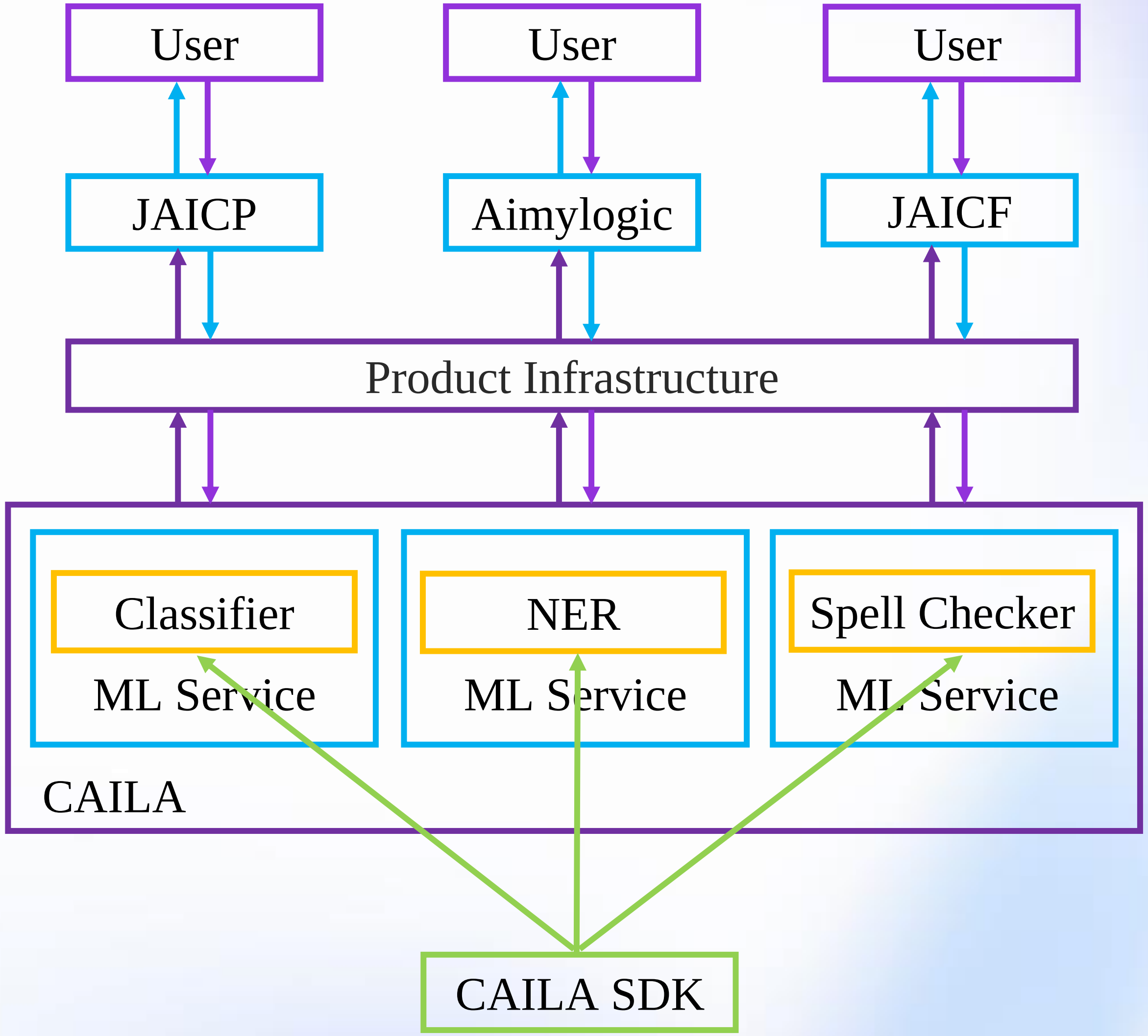
- Conversational AI Linguistic Assistant (CAILA) — internal NLP provider for company products
- Monolith service containing variety of solutions
  - Wasteful resource consumption
  - Hard to scale
  - Troubles with integrating new solutions both in NLP service and in products



CAILA 2.0: from service to platform

Just AI

- Now CAILA is a platform for creating, hosting and managing ML services
- One NLP tool — one microservice
- One public SDK for all services
  - interfaces, base classes and mixins
  - input and output data types, type checking
  - interactions with CAILA, storages, loggers
- Service languages are Python and Java
- User models can be added to CAILA via SDK
- Integration with projects in Just AI tools



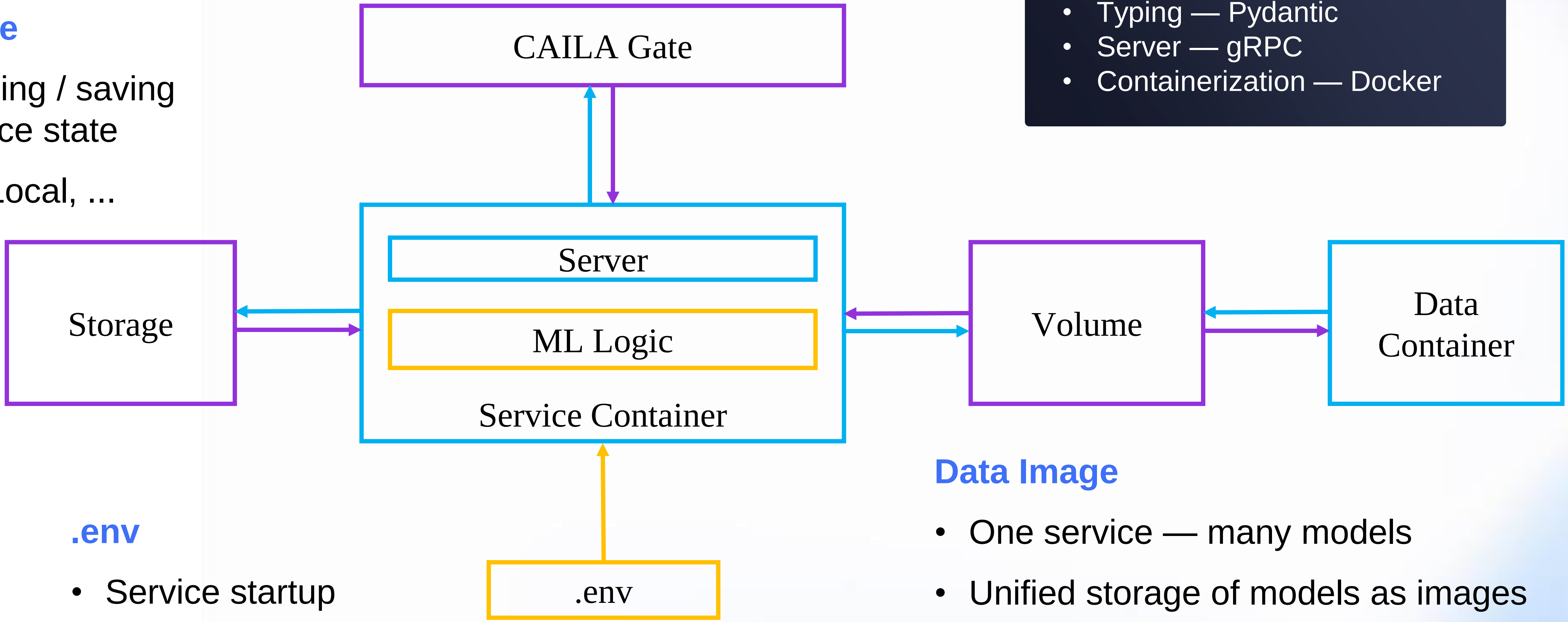
General microservice diagram

Just AI

- Typing — Pydantic
- Server — gRPC
- Containerization — Docker

Storage

- Loading / saving service state
- S3, Local, ...



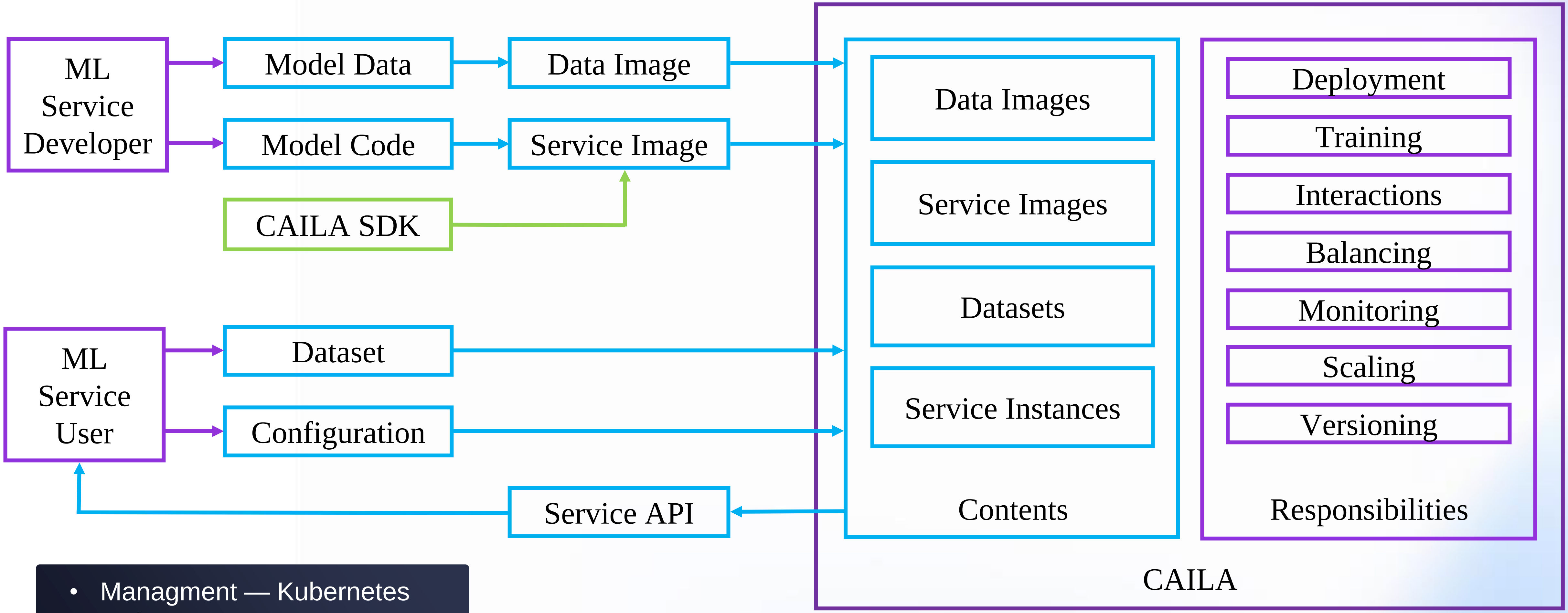
.env

- Service startup parameters
- Storage credentials

Data Image

- One service — many models
- Unified storage of models as images
- Mount in Data Container at startup

Platform overview diagram

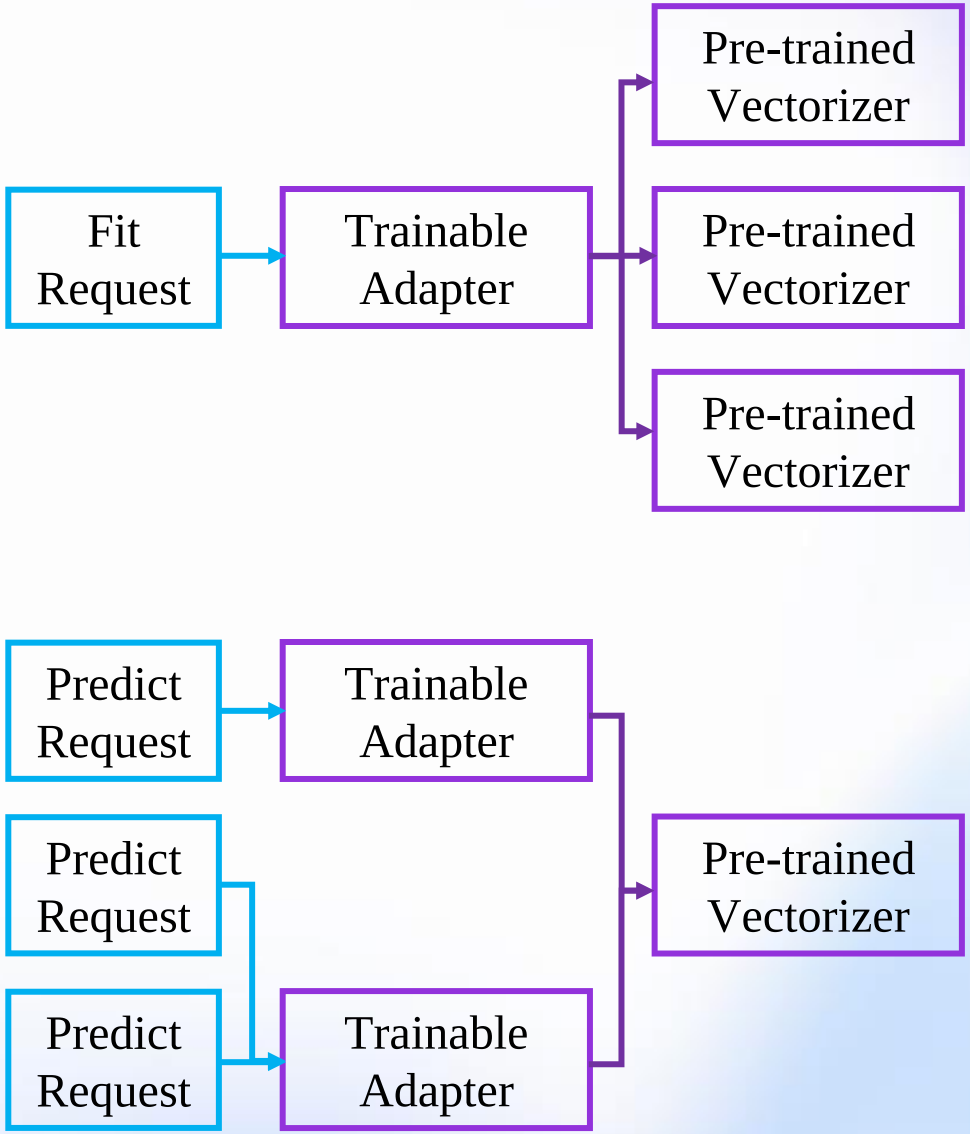


- Managment — Kubernetes
- Main storage — S3
- Main DB — Postgres
- Monitoring — Prometheus

New classifier in CAILA 2.0

Just AI

- Services can communicate with each other through CAILA Gate
- Classifier = vectorizer + adapter services
- Adapter can be ML or KNN (we use NMSLIB)
- Parallelization of heavy fit requests across vectorizers
- Dynamic batching of predict requests
- Vectors caching in storage during training



How do we use the new platform

Just AI

- Build our ML services
  - We transfer all our NLP services to CAILA
  - Integration of Just AI voice technologies is next
- Use as ML provider for our products
  - Just AI Conversational Cloud products are switching to work with CAILA services
  - The new Transformer calssifier is available in JAICP (upon request for now)
- Open CAILA to external users
  - Cloud platform with monetization for ML developers and customers
  - On-premise installations for corporate clients

Just AI

Thank you for your attention!

CAILA

app.caila.io

Murat Apishev

NLP Team Lead, Just AI

TG: @MelLain

Российский фонд развития информационных технологий